FAIR Universe 2024: Higgs ML Uncertainty Challenge

Wahid Bhimji[1], *Paolo* Calafiura[1], *Ragansu* Chakkappai[2,6], *Po-Wen* Chang[1], *Yuan-Tang* Chou[3], *Sascha* Diefenbacher[1], *Jordan* Dudley[4,1], *Steven* Farrell[1], *Aishik* Ghosh[5,1], *Isabelle* Guyon[6], *Chris* Harris[1], *Shih-Chieh* Hsu[3], *Elham E* Khoda[7,3,1], *Rémy* Lyscar[2], *Alexandre* Michon[2], *Benjamin* Nachman[1], *Peter* Nugent[1], *Mathis* Reymond[8], *David* Rousseau[2,6], *Benjamin* Sluijter[9,1], *Benjamin* Thorne[1], *Ihsan* Ullah[6], and *Yulei* Zhang[3]

[1]Lawrence Berkeley National Laboratory, USA
[2]Université Paris-Saclay, CNRS/IN2P3, IJCLab, France
[3]University of Washington, Seattle, USA
[4]University of California, Berkeley, USA
[5]University of California, Irvine, USA
[6]ChaLearn, USA
[7]University of California, San Diego, USA
[8]Université Paris-Saclay, France
[9]Universiteit Leiden, Netherlands

Abstract. The HiggsML Uncertainty Challenge is a machine learning competition aimed at improving uncertainty-aware AI techniques in high-energy physics. Part of the FAIR Universe initiative, focuses on estimating the Higgs boson signal strength while accounting for systematic uncertainties affecting collider experiments. Unlike traditional classification tasks, participants must construct confidence intervals that properly cover systematic distortions. The HiggsML Uncertainty Challenge establishes a benchmark for uncertainty-aware AI, with applications in high-energy physics and beyond. The competition is hosted on Codabench, an open AI benchmarking platform, and uses high-performance computing resources at NERSC Perlmutter for scalable and reproducible model evaluation. The dataset and evaluation framework will remain publicly available for continued research.

1 Introduction and Overview of FAIR Universe

1.1 Motivation and Background

The application of artificial intelligence (AI) and machine learning (ML) in high-energy physics (HEP) has led to significant advancements in data analysis, classification, and anomaly detection. However, a major challenge remains in quantifying and mitigating systematic uncertainties, which arise from detector effects, theoretical modeling, and background mismodeling. In many fields of science and industry, including HEP, these uncertainties, often referred to as epistemic uncertainties, are a critical bottleneck in precision measurements and discovery potential [1].

The FAIR Universe project aims to address these challenges by providing an open, large-scale AI ecosystem that enables dataset sharing, training of large models, and benchmarking

of advanced AI/ML techniques. The project addresses key challenges in measuring and minimizing systematic uncertainties, with direct relevance to applications in cosmology, climate science, and other domains. It builds upon previous AI-driven HEP challenges, including the Higgs Boson Machine Learning Challenge (HiggsML) [2], TrackML Challenge [3, 4], and LHC Olympics [5].

1.2 The FAIR Universe Platform

FAIR Universe adheres to the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) to facilitate reproducibility and open research. A core aspect of the project is leveraging high-performance computing (HPC) resources, particularly the Perlmutter system at NERSC [6], which features over 7,000 NVIDIA A100 GPUs. FAIR Universe integrates with Codabench, an open-source AI benchmarking platform designed for large-scale, transparent, and reproducible challenges [7].

The project involves a broad collaboration of institutions, including Lawrence Berkeley National Laboratory, the University of Washington, Université Paris-Saclay (IJCLab-Orsay), the University of California, Irvine, the University of California, San Diego, and Universiteit Leiden. By fostering an interdisciplinary approach, FAIR Universe aims to create benchmarks for systematic uncertainty-aware ML models and drive innovation in uncertainty quantification.

2 The HiggsML Uncertainty Challenge

High-energy physics (HEP) relies on statistical analysis of aggregated observations, where systematic uncertainties from detector effects, theoretical modeling, and background mismodeling can bias measurements of key quantities such as the Higgs boson signal strength (μ). As machine learning (ML) becomes increasingly central to HEP analyses, the need for uncertainty-aware methods has grown.

Advanced techniques for incorporating systematic uncertainties into ML models include approaches that explicitly depend on nuisance parameters [10–13], those that aim to be insensitive to nuisance parameters [14–23], methods that incorporate downstream test statistics in initial training [24–28], and Bayesian neural networks for uncertainty estimation [29–31].

The **HiggsML Uncertainty Challenge**, part of the FAIR Universe project, advances uncertainty-aware AI in high-energy physics by requiring models to estimate the Higgs boson signal strength (μ) with statistically valid confidence intervals (CIs) while accounting for systematic uncertainties. The competition is part of **NeurIPS 2024** [8] and hosted on **Codabench** [9].

The challenge consists of two phases. In Phase 1, participants train models on a labeled dataset with systematic uncertainties and refine them iteratively based on daily leaderboard feedback. In Phase 2, the best models are tested on a hidden dataset with unknown biases, with final rankings based on the accuracy and robustness of CI estimation. The competition workflow is shown in Figure 1.

The challenge introduces six systematic distortions, requiring models to estimate μ accurately, provide reliable CIs, and remain robust to distribution shifts. The HiggsML Uncertainty Challenge thus serves as a benchmark to compare uncertainty-aware ML approaches on a realistic dataset, ensuring their reliability for future physics analyses.

Beyond high-energy physics, this benchmark has applications in fields where systematic uncertainties impact precision, such as cosmology, climate science, and medical imaging.

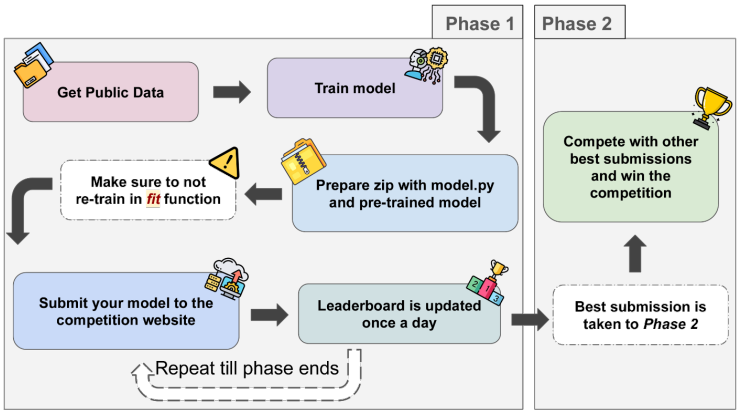

Figure 1. Workflow of the HiggsML Uncertainty Challenge.

3 Dataset and Systematic Uncertainties

3.1 Dataset Overview

The dataset used in the HiggsML Uncertainty Challenge is a simulated high-energy physics dataset representing proton-proton collisions at the Large Hadron Collider (LHC). It is generated using **Pythia 8.2** for event simulation [32] and **Delphes 3.5.0** for fast detector simulation [33]. The dataset is designed to reflect the kinematic distributions observed in the ATLAS experiment [34], making it a valuable benchmark for machine learning models in high-energy physics.

Each collision event is represented by **28 features**, which include both raw kinematic measurements and high-level physics-inspired variables. These features characterize the reconstructed particles in an event and are essential for signal-background discrimination. The dataset is divided into two main categories: Signals are the events corresponding to Higgs boson production, specifically the decay process $H \to \tau\tau$, where one tau decays leptonically, and the other decays hadronically. Backgrounds are the events arising from other Standard Model processes that mimic the kinematic signature of Higgs boson decays, including $Z \to \tau\tau$, top-antitop ($t\bar{t}$), and diboson (VV) production. To ensure realistic scaling and balanced representation across processes, the "LHC Equivalent Events" column reports the expected yields for 10 fb^{-1} at $\sqrt{s} = 13$ TeV, estimated using known production cross-sections and branching ratios. This ensures sufficient statistics while avoiding large event weights, thereby reducing Monte Carlo uncertainties. A summary of the dataset composition is presented in Table 1.

Table 1. Composition of the HiggsML Uncertainty Challenge dataset. The number of events corresponds to the total generated sample, while the LHC-equivalent column estimates the expected event counts in a 10 fb^{-1} dataset.

Process	Events Generated	LHC Equivalent Events	Category
Higgs Boson ($H \to \tau\tau$)	52M	1015	**Signal**
Z Boson ($Z \to \tau\tau$)	222M	1002K	**Background**
Diboson (VV)	2.1M	3.8K	**Background**
Top-Antitop ($t\bar{t}$)	12M	44K	**Background**

The dataset is **publicly available** for participants during the competition and will remain accessible as a benchmark dataset after the challenge.

3.2 Systematic Uncertainties and Nuisance Parameters

A key aspect of the HiggsML Uncertainty Challenge is the explicit incorporation of systematic uncertainties, which introduce biases in the test dataset. These uncertainties arise from detector calibration, theoretical modeling, and background estimation. Understanding and mitigating them is essential for precision measurements in high-energy physics.

To simulate these effects, the dataset includes six parameterized systematic distortions, known as nuisance parameters. These distortions modify event features, creating realistic variations between training and test datasets.

The systematic uncertainties modeled in this challenge are the tau and jet energy scales, soft missing transverse energy shifts, and normalization uncertainties for the overall background, diboson background, and top-antitop background. These distortions introduce nontrivial biases, requiring models to account for uncertainty while estimating the Higgs boson signal strength. Participants must develop methods to ensure robustness against systematic variations.

Figure 2 illustrates the impact of systematic uncertainties on key event features. The distributions of nominal events (blue) and systematically shifted events (red) are overlaid, showing how different nuisance parameters distort kinematic variables such as transverse momentum (**PRI_had_pt**), missing energy (**PRI_met**), and invariant masses (**DER_mass_vis, DER_mass_transverse_met_lep**). These shifts highlight the importance of designing models that can adapt to systematic variations in the data.

3.3 Biasing Script and Pseudo-Experiments

To model systematic uncertainties, the competition provides a biasing script that generates test datasets with arbitrary nuisance parameter settings. This script introduces controlled distortions in event kinematics and simulates realistic deviations between training and test datasets.

A key component is the use of pseudo-experiments, which replicate independent statistical realizations of the data under different systematic conditions. Each pseudo-experiment represents a dataset corresponding to a given signal strength μ while randomly varying nuisance parameters. Participants must analyze distributions of μ across multiple pseudo-experiments to ensure their models provide statistically valid confidence intervals. This setup mimics real-world experimental conditions, where systematic effects fluctuate between data-taking periods.

4 Evaluation Metrics

The challenge aims to ensure that models not only estimate the Higgs boson signal strength (μ) accurately but also provide a reliable 68.27% confidence interval (CI) that accounts for systematic uncertainties. Evaluation is based on two criteria: precision, which favors narrower confidence intervals, and coverage, which measures the fraction of test cases where μ_{true} falls within the reported interval. To balance these aspects, the challenge introduces a novel metric, the Coverage Score, which quantifies model performance by considering both precision and coverage.

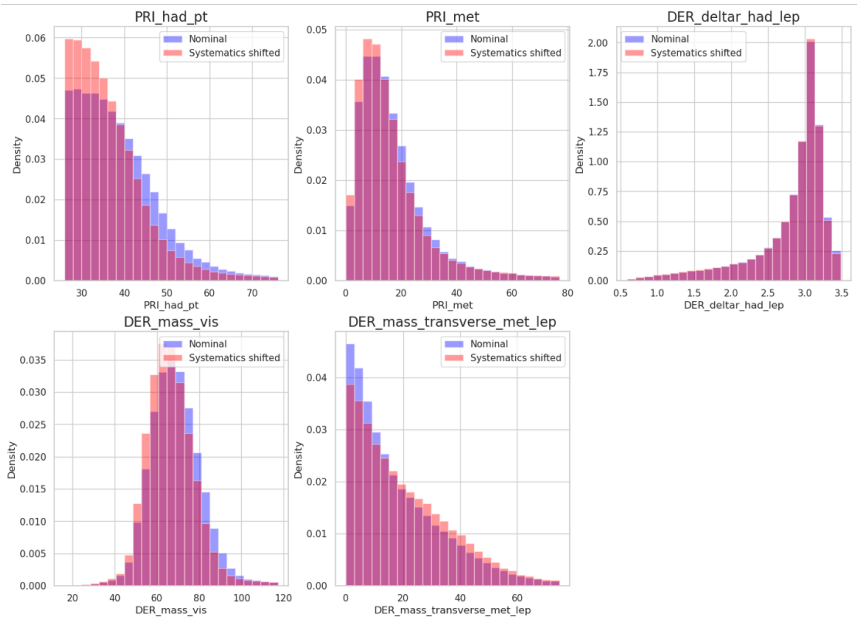

Figure 2. Impact of systematic uncertainties on key event features in the HiggsML dataset. Each subplot compares the nominal feature distribution (blue) to its systematically shifted counterpart (red), illustrating how different nuisance parameters affect the observed data. These variations must be accounted for in model predictions to ensure robustness.

4.1 Coverage Score Calculation

The Coverage Score consists of two components:

- **Average Interval Width** (w): Measures the typical size of the confidence interval across test samples.

$$w = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |\mu_{84,i} - \mu_{16,i}| \tag{1}$$

  where $\mu_{84,i}$ and $\mu_{16,i}$ represent the upper and lower bounds of the confidence interval in each pseudo-experiment.

- **Coverage** (c): The fraction of test cases where the true signal strength μ_{true} is within the estimated confidence interval.

$$c = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} 1 \quad \text{if} \quad \mu_{\text{true},i} \in [\mu_{16,i}, \mu_{84,i}] \tag{2}$$

The ideal coverage value is 0.6827 (corresponding to the 68.27% confidence interval), but models may overestimate or underestimate uncertainties. To address this, a penalization function $f(c)$ is applied:

$$f(c) = \begin{cases} 1 & \text{if } c \in [0.6827 - 2\sigma_{68}, 0.6827 + 2\sigma_{68}] \\ 1 + \left| \frac{c - (0.6827 - 2\sigma_{68})}{\sigma_{68}} \right|^4 & \text{if } c < 0.6827 - 2\sigma_{68} \\ 1 + \left| \frac{c - (0.6827 + 2\sigma_{68})}{\sigma_{68}} \right|^3 & \text{if } c > 0.6827 + 2\sigma_{68} \end{cases} \tag{3}$$

where $\sigma_{68} = \sqrt{(1-0.6827)0.6827/N_{\text{test}}}$ represents the statistical uncertainty on the coverage estimation. This function penalizes deviations from the target coverage, applying a stronger penalty for underestimation and a milder penalty for overestimation.

The final Coverage Score is calculated as:

$$\text{score} = -\ln((w+\epsilon)f(c)) \tag{4}$$

where $\epsilon = 10^{-2}$ is a regularization term preventing unrealistic confidence intervals. Higher scores indicate better performance.

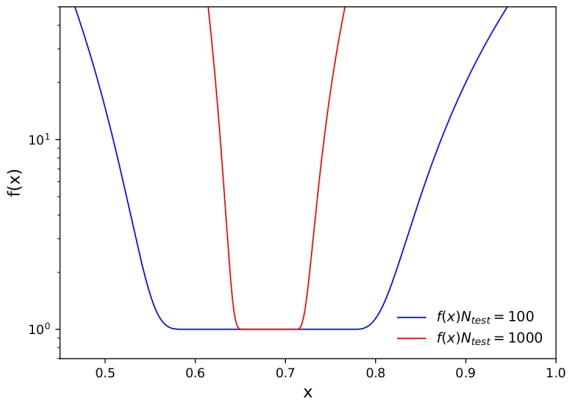

Figure 3. Illustration of the Coverage Score metric. The penalty function ensures that models achieving coverage close to 68.27% are rewarded, while those significantly under- or over-covering are penalized.

4.2 Baseline Model

A baseline model is provided in the challenge repository as a reference for participants. It includes a Gradient Boosted Decision Tree (XGBoost) [35] classifier for signal-background separation, a likelihood-based signal strength estimator, and a simple Bayesian approach to model systematic uncertainties. While not fully accounting for systematic distortions, it serves as a foundation for developing more robust models.

4.3 Computing Environment

The challenge is hosted on Codabench [7] and utilizes NERSC Perlmutter [6], a supercomputer with over 7000 NVIDIA A100 GPUs, for model evaluation. Participants train their models locally or on external resources of their choice, and submit prediction code for evaluation. Each submission is allocated a maximum of 2 hours, with up to 10 seconds per pseudo-experiment. Participants may submit up to 100 times per day, with five evaluations processed daily.

5 Conclusion and Outlook

The **HiggsML Uncertainty Challenge** establishes a benchmark for **uncertainty-aware machine learning** in high-energy physics. By incorporating systematic uncertainties, it moves

beyond classification to robust confidence interval estimation for the Higgs boson signal strength (μ). The large-scale dataset and rigorous evaluation metric ensure fair assessment of model performance.

This challenge bridges machine learning and physics inference, with broader applications in fields where systematic uncertainties impact precision. Future directions include expanding the dataset, integrating real experimental data, and developing advanced uncertainty quantification techniques. The dataset and framework will remain publicly available to support ongoing research in uncertainty-aware AI.

References

[1] C. Chen et al., Uncertainty-aware Machine Learning for High Energy Physics, *SciPost Phys. Codebases* **7** (2022). https://doi.org/10.21468/SciPostPhys.7.1.004

[2] G. Cowan et al., The Higgs Boson Machine Learning Challenge, *Proceedings of Machine Learning Research*, **42**, 19 (2014). https://proceedings.mlr.press/v42/cowa14.html

[3] S. Amrouche et al., The TrackML Challenge: Accuracy Phase, *Comput. Softw. Big Sci.* **3**, 1 (2019). https://doi.org/10.1007/s41781-019-0028-7

[4] D. Rousseau et al., The TrackML Challenge: Throughput Phase, *Comput. Softw. Big Sci.* **4**, 6 (2020). https://doi.org/10.1007/s41781-020-00046-5

[5] G. Kasieczka et al., The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics, *SciPost Phys. Codebases* **5** (2021). https://doi.org/10.21468/SciPostPhys.5.1.003

[6] NERSC Perlmutter System, https://www.nersc.gov/systems/perlmutter/

[7] I. Guyon et al., Codabench: An Open AI Benchmarking Platform, *NeurIPS Competitions Track* (2021). https://neurips.cc/Conferences/2021/CompetitionTrack

[8] NeurIPS 2024 Competition Track, https://neurips.cc/Conferences/2024/CompetitionTrack

[9] FAIR Universe, HiggsML Uncertainty Challenge, https://www.codabench.org/competitions/2977/

[10] B. Nachman et al., Adversarially Learned Inference, *Phys. Rev. D* **101**, 083011 (2020). https://doi.org/10.1103/PhysRevD.101.083011

[11] K. Cranmer, J. Pavez, and G. Louppe, Approximating Likelihood Ratios with Calibrated Discriminative Classifiers, arXiv:1506.02169.

[12] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, How to Inject Background Knowledge into a Neural Network, arXiv:1604.05400.

[13] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Constraining Effective Field Theories with Machine Learning, *Phys. Rev. Lett.* **124**, 011801 (2020). https://doi.org/10.1103/PhysRevLett.124.011801

[14] A. Blance, M. Spannowsky, and P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches, *JHEP* **10**, 047 (2019). https://doi.org/10.1007/JHEP10(2019)047

[15] C. Englert, P. Galler, P. Harris, and M. Spannowsky, Machine Learning Uncertainties with Adversarial Neural Networks, *Eur. Phys. J. C* **79**, 4 (2019). https://doi.org/10.1140/epjc/s10052-018-6511-8

[16] L. Bradshaw, R. Mishra, A. Mitridate, and B. Ostdiek, Mass Agnostic Jet Taggers, *SciPost Phys.* **8**, 011 (2020). https://doi.org/10.21468/SciPostPhys.8.1.011

[17] ATLAS Collaboration, Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS, ATL-PHYS-PUB-2018-014 (2018). http://cds.cern.ch/record/2630973

[18] G. Kasieczka and D. Shih, DisCo Fever: Robust Networks Through Distance Correlation, *Phys. Rev. Lett.* **125**, 122001 (2020). https://doi.org/10.1103/PhysRevLett.125.122001

[19] S. Wunsch, S. Jörger, R. Wolf, and G. Quast, Reducing the dependence of the neural network function to systematic uncertainties in the input space, *Comput. Softw. Big Sci.* **4**, 5 (2020). https://doi.org/10.1007/s41781-020-00037-9

[20] A. Rogozhnikov et al., New approaches for boosting to uniformity, *JINST* **10**, T03002 (2015). https://doi.org/10.1088/1748-0221/10/03/T03002

[21] CMS Collaboration, A deep neural network to search for new long-lived particles decaying to jets, *Machine Learning: Science and Technology* **1**, 035012 (2020). https://doi.org/10.1088/2632-2153/ab9023

[22] V. Estrade, C. Germain, I. Guyon, and D. Rousseau, Systematic aware learning - A case study in High Energy Physics, *EPJ Web Conf.* **214**, 06024 (2019). https://doi.org/10.1051/epjconf/201921406024

[23] A. Ghosh and B. Nachman, A cautionary tale of decorrelating theory uncertainties, *Eur. Phys. J. C* **82**, 46 (2022). https://doi.org/10.1140/epjc/s10052-022-10012-w

[24] S. Wunsch et al., Optimal Statistical Inference in the Presence of Systematic Uncertainties Using Neural Network Optimization Based on Binned Poisson Likelihoods with Nuisance Parameters, *Comput. Softw. Big Sci.* **5**, 4 (2021). https://doi.org/10.1007/s41781-020-00049-5

[25] A. Elwood, D. Krücker, and M. Shchedrolosiev, Direct optimization of the discovery significance in machine learning for new physics searches, *J. Phys. Conf. Ser.* **1525**, 012110 (2020).

[26] N. Simpson and L. Heinrich, neos: End-to-End-Optimised Summary Statistics for High Energy Physics, *J. Phys. Conf. Ser.* **2438**, 012105 (2023). https://doi.org/10.1088/1742-6596/2438/1/012105

[27] P. Feichtinger et al., Punzi-loss: a non-differentiable metric approximation for sensitivity optimisation, *Eur. Phys. J. C* **82**, 121 (2022). https://doi.org/10.1140/epjc/s10052-022-10070-0

[28] L. Layer, T. Dorigo, and G. Strong, Application of Inferno to a Top Pair Cross Section Measurement with CMS Open Data, arXiv:2301.10358.

[29] G. Kasieczka et al., Per-Object Systematics using Deep-Learned Calibration, *SciPost Phys.* **9**, 089 (2020). https://doi.org/10.21468/SciPostPhys.9.6.089

[30] J. Y. Araz and M. Spannowsky, Combine and Conquer: Event Reconstruction with Bayesian Ensemble Neural Networks, *JHEP* **04**, 296 (2021). https://doi.org/10.1007/JHEP04(2021)296

[31] M. Bellagente, M. Haußmann, M. Luchmann, and T. Plehn, Understanding Event-Generation Networks via Uncertainties, arXiv:2104.04543.

[32] T. Sjöstrand et al., An Introduction to PYTHIA 8.2, *Comput. Phys. Commun.* **191**, 159-177 (2015). https://doi.org/10.1016/j.cpc.2015.01.024

[33] J. de Favereau et al. [Delphes Collaboration], DELPHES 3: A modular framework for fast simulation of a generic collider experiment, *JHEP* **02**, 057 (2014). https://doi.org/10.1007/JHEP02(2014)057

[34] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, *JINST* **3**, S08003 (2008). https://doi.org/10.1088/1748-0221/3/08/S08003

[35] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, *CoRR*, **abs/1603.02754** (2016). http://arxiv.org/abs/1603.02754